Cross-lingual Search Based on Concepts and Meaning

BABELSTREET

Executive Summary

Cross-lingual search is the process of querying in one language to find relevant documents in other languages. Until recently, machine translation has been the primary method of searching across languages either by translating search queries into other languages or translating searchable records into English. However, machine translation, or for that matter human translation, loses valuable nuances and meaning present in the original text.

This white paper explores an approach that delivers better accuracy based on semantics (meaning), not translation. Semantic search (also known as concept search) goes beyond finding keywords to retrieving ideas suggested by the keywords.

In part 1, we compare the traditional translation-based approach with a newer approach that uses semantic similarity through text embeddings—a way to represent words in natural language processing tasks that encodes the meaning of words as mathematical vectors.

In part 2, we look at implementing semantic search as we discuss:

- How to retrofit an existing keyword search engine to add cross-lingual and fuzzy search
- Ways to overcome issues of speed, especially when searching very large data sets
- A specific use case — targeted topic and event extraction
- The special case of cross-lingual name matching

> Today's searchers expect answers within seconds, and that is a challenge for semantic search which relies on relatively "slow" vector comparisons.

Part 1: Comparing Technologies

Translation-based cross-lingual information retrieval

Multiple situations require cross-lingual search, such as:

- intelligence analysts mining open source data (OSINT)
- lawyers doing e-discovery to find relevant documents for a case
- patent lawyers researching technical documents

To mimic cross-lingual search, people use online translation platforms to laboriously find the equivalent terms and then re-execute the query multiple times in different languages. The second step assumes that you have a search engine equipped with the multilingual processing smarts to search in languages other than English.

Until recently, the commercial search industry hasn't seen much demand for cross-lingual search. Search has always been monolingual and very English-centric. Multilingual search was interesting only as far as being able to search in English plus another language.

Challenges of query translation

Cross-lingual search was mainly a research challenge for labs and universities, which would take one of two approaches. One was to machine translate the query (most often from English) into the target language(s).

Translate query into multiple languages and search in each language

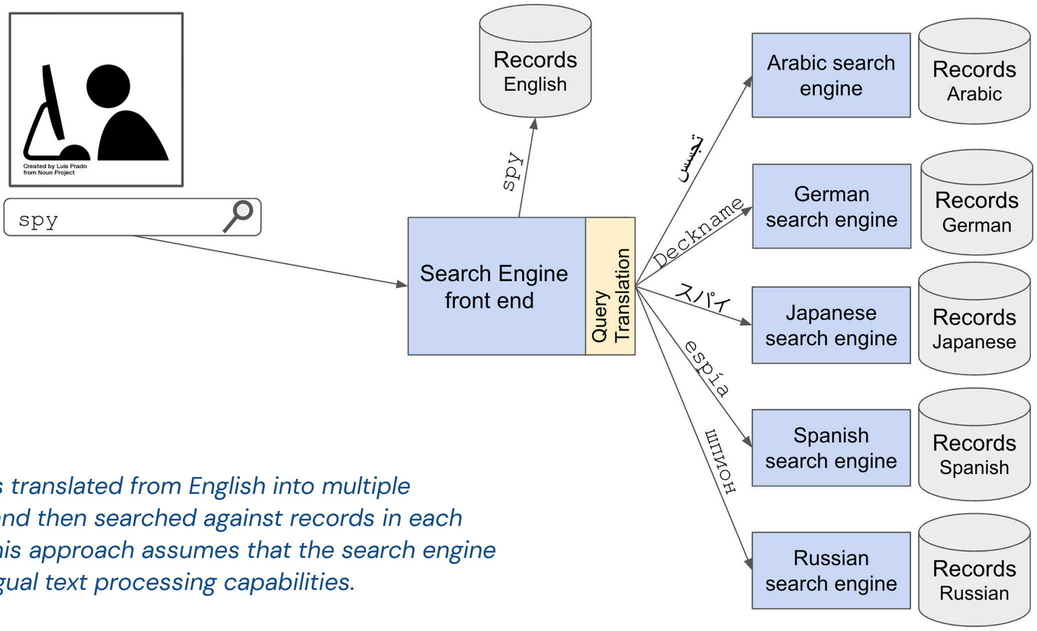
However, queries of one to three words are often too short to provide enough context to produce a good machine translation. Words can be used in multiple contexts and have multiple meanings. The query “tree program” can be interpreted as “a program to increase tree canopy,” “a way to populate Minecraft with trees,” or “a computer program to list a directory structure.”

There have been numerous strategies to improve machine translation for search. Here is a small sampling:

- Combine dictionary lookup with ontologies[1] to reconfirm the correct set of keywords, or use a thesaurus to broaden the search by adding more and less specific terms related to the keyword[2]
- Apply special translation algorithms for a particular domain, such as technical documents[3]
- Use “a translation lattice containing all possible translations with their respective probabilities of accuracy”[4]

Machine translation and loss of fidelity

A second approach is to machine translate all the non-English searchable records to English, which has much more context, but is still not enough.



The query is translated from English into multiple languages, and then searched against records in each language. This approach assumes that the search engine has multilingual text processing capabilities.

[1] An ontology for a specific domain defines a set of concepts and categories that shows the properties and relationships between them.

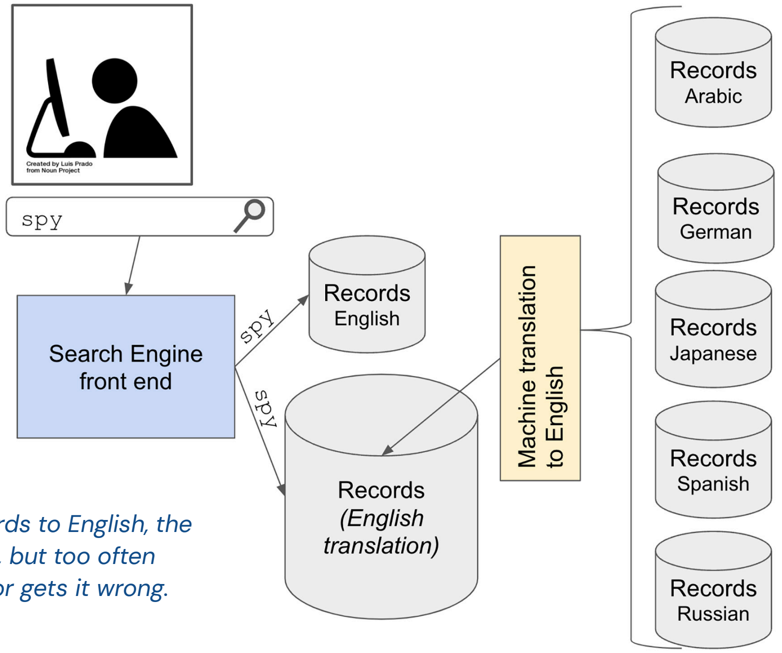
[2] Reddy, Mallamma, et al “Cross Lingual Information Retrieval Using Search Engine and Data Mining” ACEEE Int. J. on Information Technology, Vol. 01, No. 02, Sep 2011 https://www.researchgate.net/publication/228395995_Cross_Lingual_Information_Retrieval_Using_Search_Engine_and_Data_Mining

[3] Strategies specific to how to translate technical terms between English and Japanese for cross-lingual search of technical documents are discussed in Fujii, Atsushi & Ishikawa, Tetsuya (2002) “Cross-Language Information Retrieval for Technical Documents” University of Library and Information Science http://aclweb.org/anthology/W99-0605

[4] Tantillo, Ariana “A cross-language search engine enables English monolingual researchers to find relevant foreign-language documents” Massachusetts Institute of Technology, May 30, 2016 https://phys.org/news/2016-05-cross-language-enables-english-monolingual-relevant.html

Translate all records into English and search in English

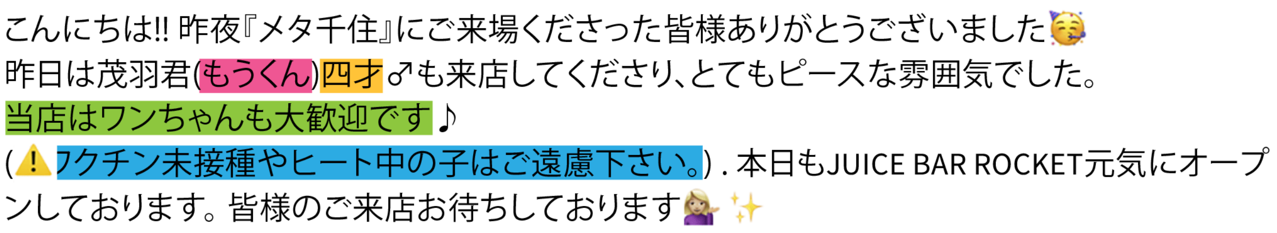
The challenge comes down to machine translation in general. Documents provide a lot more context, but the translations can still be laughably wrong, especially between linguistically distant languages. (Douglas Hofstadter takes a fascinating look at what makes machine translation so difficult in "The Shallowness of Google Translate: The program uses state-of-the-art AI techniques, but simple tests show that it's a long way from real understanding."[5])



By translating all the searchable records to English, the problem is reduced to English search, but too often the machine translation is off target or gets it wrong.

Example 1

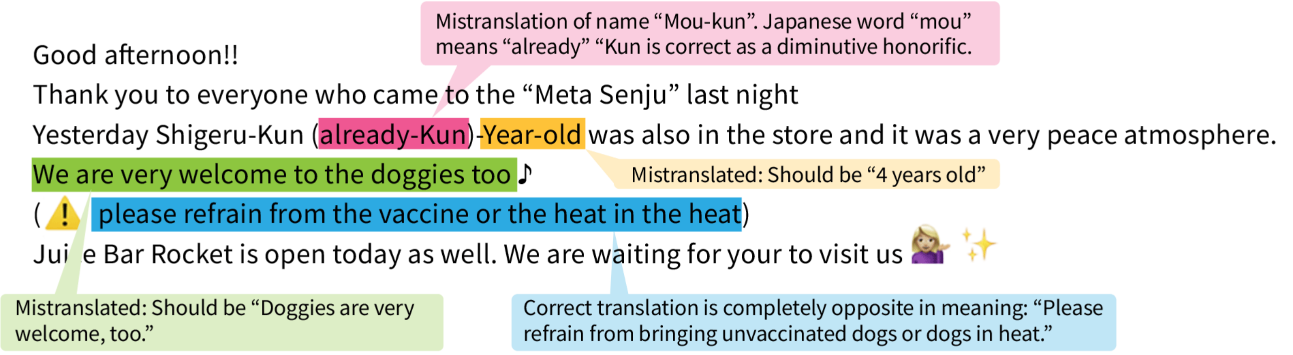
Here is a machine-translated social media post in Japanese:[6]



こんにちは!! 昨夜『メタ千住』にご来場くださった皆様ありがとうございました🥳
昨日は茂羽君(もうくん)四才♂も来店してくださり、とてもピースな雰囲気でした。
当店はワンちゃんも大歓迎です♪
(⚠️ワクチン未接種やヒート中の子はご遠慮下さい。) . 本日もJUICE BAR ROCKET元気にオープンしております。皆様のご来店お待ちしております💁‍♀️ ✨

[5] Hofstadter, Doug "The Shallowness of Google Translate: The program uses state-of-the-art AI techniques, but simple tests show that it's a long way from real understanding." The Atlantic, January 30, 2018. https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

[6] Source: Juice Bar Rocket social media post https://www.yooying.com/p/1996282483286416069_3204056243. Machine translated with Google Translate in March 2019. Retrieved and machine translated with Google Translate in March 2019.



Mistranslation of name “Mou-kun”. Japanese word “mou” means “already” “Kun is correct as a diminutive honorific.

Good afternoon!!
Thank you to everyone who came to the “Meta Senju” last night
Yesterday Shigeru-Kun (already-Kun)-Year-old was also in the store and it was a very peace atmosphere.
We are very welcome to the doggies too♪
(⚠️ please refrain from the vaccine or the heat in the heat)
Jui e Bar Rocket is open today as well. We are waiting for your to visit us 💁 ✨

Mistranslated: Should be “4 years old”

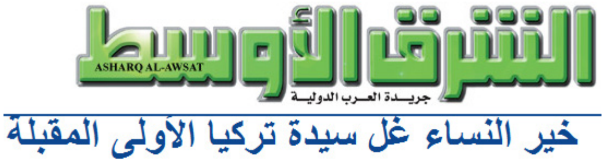
Mistranslated: Should be “Doggies are very welcome, too.”

Correct translation is completely opposite in meaning: “Please refrain from bringing unvaccinated dogs or dogs in heat.”

Example 2

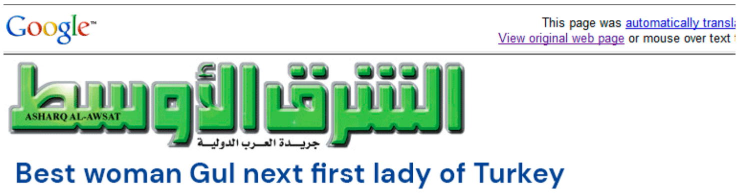
In the following example, a name is mistakenly translated instead of transliterated because it is not recognized by the system as a proper noun.

Original Arabic news article headline:



الشرق الأوسط
ASHARQ AL-AWSAT
جريدة العرب الدولية
خير النساء غل سيدة تركيا الأولى المقبلة

Machine translated to English:



Google
This page was automatically transl
View original web page or mouse over text
الشرق الأوسط
ASHARQ AL-AWSAT
جريدة العرب الدولية
Best woman Gul next first lady of Turkey

When Hayrunnisa Gul was announced as the 11th first lady of Turkey, her given name in an Arabic news article was literally translated to English as “Best woman” instead of being transliterated as Hayrunnisa.

New semantics-based cross-lingual information retrieval

The semantics-based approach uses text embeddings, which convert the meaning of words into mathematical vectors.

The idea of expressing the meaning of words as mathematical vectors appeared in the 1960s with Cornell University’s SMART system, and most text information retrieval models since the ’80s have been based on vector models of text. However, it was only starting around 2013 that the task of creating word vectors became much more scalable and practical with the introduction of new algorithms. A dense/distributed vector representation can encode word semantics in ways that more naturally generalize to words which have not been seen before. Text embeddings are based on the idea that words and phrases whose vectors are close in vector space will also be close in meaning, whether they are written in the same or a different language.

The technology

The vectorization of text is based on training models on a huge collection of raw documents (on the relevant topics and domains) for each language and assigning vectors for each language’s “space,” such that the words with similar meanings (between languages) have similar values. The degree to which two words are similar is expressed by a matching score that reflects the similarity of their corresponding vectors.

Here is an example of the distance between the word “spy” and its semantically similar terms in Spanish, German, and Japanese. *NOTE: The same word compared to itself scores a perfect 1.0.*

Spanish

```
{"term":"espía","similarity":0.61295485},
{"term":"cia","similarity":0.46201307},
{"term":"desertor","similarity":0.42849663},
{"term":"cómplice","similarity":0.36646274},
{"term":"subrepticiamente","similarity":0.36629659}
```

German

```
{"term":"Deckname","similarity":0.51391315},
{"term":"GRU","similarity":0.50809389},
{"term":"Spion","similarity":0.50051737},
{"term":"KGB","similarity":0.49981388},
{"term":"Informant","similarity":0.48774603},
```

Japanese

```
{"term":"スパイ","similarity":0.5544399},
{"term":"諜報","similarity":0.46903181},
{"term":"MI6","similarity":0.46344957},
{"term":"殺し屋","similarity":0.41098994},
{"term":"正体","similarity":0.40109193},
```

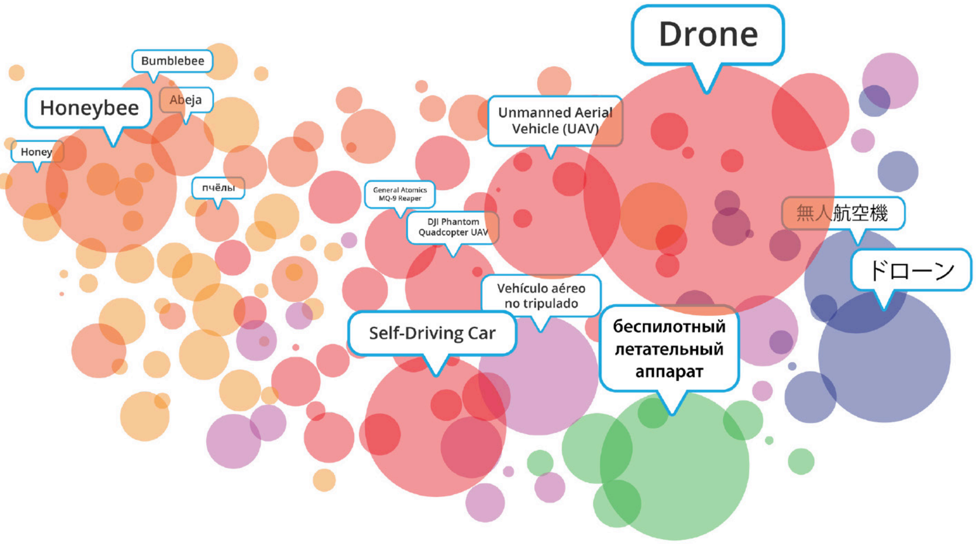
As with any NLP technology, text embeddings will produce better results if the training corpus matches the domain of the search engine. If there are not enough articles about espionage, or enough contexts in which “spy” appears, then it will be difficult to identify similar terms related to “spy.”

So far, we have talked about one vector per word. It’s also possible to create vectors for a phrase, a sentence, or even an entire document. The linear nature of the individual word vectors enables them to be aggregated in a meaningful way to represent the meaning of a text segment. This approach is similar to a “bag of words” in which the position of the word in the text is not taken into consideration, merely the fact that the words appear in the same document is enough. However, there are much more comprehensive approaches to represent a text segment using vectors that take the location of the words into consideration.

Use case: Concept search

Concept search is a good method for broadening a search to include fuzzy matches. Here, fuzziness also enables the searcher to say “find me more documents like this one” when an entire document is the search query, or “find documents that are relevant, even if they don’t contain the actual keywords.”

For example, a fuzzy search on “flying drones” using text embeddings may return articles about “unmanned aerial vehicle (UAV)” or specific models of UAVs, such as “DJI Phantom quadcopter” and “General Atomics MQ-9 Reaper.” Fuzzy matching means searchers unfamiliar with the term “UAV” will still be able to find highly relevant articles for “flying drones.” Fuzzy search is like expanding circles around meaning, so another level might include “autonomous robots.”


Bumblebee
Honeybee
Abeja
Honey
пчёлы
General Atomics MQ-9 Reaper
DJI Phantom Quadcopter UAV
Unmanned Aerial Vehicle (UAV)
Drone
無人航空機
ドローン
Vehículo aéreo no tripulado
Self-Driving Car
беспилотный летательный аппарат

Part 2: Implementation

In this section we look at the mechanics of implementing semantic search:

- How to retrofit existing keyword search to add fuzzy, cross-lingual search
- How to use semantic similarity for targeted event detection
- Challenges and solutions related to execution speed and searching large data sets

Retrofitting keyword search with cross-lingual semantic search

In existing keyword search systems, the addition of cross-lingual semantic search involves introducing text embeddings at both indexing time and query execution.

- At indexing time, calculate a text embedding value for each document or sentence of your searchable records, then store these vector values in the index
- At query time, vectorize the search keywords and run a comparison of the query's vector against the indexed vectors, then return results based on a mathematical comparison to find the "closest match"

Challenges

Getting multilingual results

English search terms will match English results before matching similar terms in other languages. So, in order to show multilingual results in a dataset with millions of records, the system must be set up to return the top results from each language that pass a certain matching threshold. Without this special handling, search results using semantic similarity will look similar to plain, monolingual keyword search.

Getting fuzzy results

Modern keyword search engines are good at finding relevant results around the given keywords. The semantic search will also return these exact word matches, so there's some configuration and experimentation required to exclude literal matches from the fuzzy search results. This way "fuzzy results" appear at the top of the results pages.

Speed and data size

Today's searchers expect answers within seconds, and that is a challenge for semantic search which relies on relatively "slow" vector comparisons. On commodity hardware, approximately 5,000 to 6,000 transactions (mathematical comparisons) per second is a reasonable expectation—up to a maximum of 25,000 transactions, if you throw more hardware at the problem.

These rates are still inadequate to search just one terabyte of data — the equivalent of 1.8 billion paragraphs of text. On a single-threaded system running 3,700 transactions per second, each query would take 74 hours. To execute searches in one second would require 1.5 million processor cores!

There exist incredibly fast and accurate specialized indexing algorithms that will parse data down to a tree and run queries in memory. In our experiments, there were speed increases on the scale of 5,000 transactions per second to 50 million transactions per second, with no degradation in accuracy. For yet more throughput, it is possible to push the system to 400 million transactions per second, although some accuracy is lost. Instead of 20 of the top 20 results being relevant, you might only get 15 out of 20, and depending on your goals, this type of trade-off may be acceptable.

More Than Entities — Finding Topics and Events

People, organizations, and locations are frequently the focus of searches. Yet there are also times when search is for events, such as a marketing department tracking the buzz around a product launch. Or a topic might be the focus, such as restricting a search to articles about crime and wrongdoing articles when performing due diligence on potential business partners. Both goals can be fulfilled with semantic search.

Event extraction is still in its early stages. A simple approach looks for a bag of words — such as buy, sell, contract — that indicate a money transaction event, with rules to capture inflected forms of those words. More sophisticated extraction can find an entire sentence indicating an event, such as "IBM sold Lotus to HCL." Semantic search offers better results than a bag of words with much less training and complexity than event extraction.

The approach using text embeddings is often called "transfer learning," as machine learning performed in one language gives a head start to training a model for other languages.

Suppose you want to look for cyberattack events. You would start by drawing up a bag of words with terms like phishing, denial of service, or hacking, for example. Then, instead of only matching against those words, you would add semantically similar words to pick up mentions such as "SQL injection," "cross-site scripting," "MitM," and "session hijacking." In our experiments, when matches were limited to those scoring 0.8 and higher — on a scale of no match (0) to perfect match (1) — accuracy of about 80% was possible. Although semantic searches will not provide human-level event extraction, they will beat a simplistic bag of words approach because they match meaning, not just literal words.

The approach using text embeddings is often called "transfer learning," as machine learning performed in one language gives a head start to training a model for other languages. How? Text embeddings link words with similar meanings between languages, thus transferring some of what is learned from a previous language to a new language.

Use case: eCommerce

Online shopping is a perfect use case for fuzzy, semantic search. Shoppers rarely know the exact name of a product they need, but they know the rough description. A raincoat might be described as "rain shell," "Gore Tex shell," or "waterproof jacket." A search engine using text embeddings would be able to match the search words to a description and a product ID number.

Cross-lingual name search

We need to mention one special case in cross-lingual search: names. Recall the earlier example about the "Best woman Gul," where Ms. Hayrunnisa Gul's given name was mistakenly literally translated as "Best woman." Semantic similarity will not help recognize when two names are the same, if they are written in two different languages or scripts. There simply cannot be enough names and context around the names in any training data to do so.

What does work is pre-extracting names of people, places, and organizations from text to a metadata field, and then doing a direct fuzzy comparison between two names *without any translation*. (See "An Overview of Fuzzy Name Matching Techniques" for a full discussion of this topic.) To distinguish between different people with similar names, entity linking (also known as entity resolution) uses the context around the name to determine which real-life person the name maps to, using a knowledge base.

Text embeddings offer better results than a "bag of words" and with much less training and complexity than relationship extraction.

Conclusion

Semantic similarity is still in its early days in terms of wide deployment in production systems of companies outside of the tech giants, but it is proven technology that is ready to replace the last generation of methods which has dominated NLP solutions to date. Semantic similarity is a smart approach that bridges cross-lingual barriers by using the technology to directly analyze native text, instead of using machine translation — and thus preserves fidelity in the analysis.

Babel Street is the trusted technology partner for the world's most advanced identity intelligence and risk operations. The Babel Street Insights platform delivers advanced AI and data analytics solutions to close the Risk-Confidence Gap.

Babel Street provides unmatched, analysis-ready data regardless of language, proactive risk identification, 360-degree insights, high-speed automation, and seamless integration into existing systems. We empower government and commercial organizations to transform high-stakes identity and risk operations into a strategic advantage.

Learn more at **babelstreet.com**.